ISSN 1561-5359. Штучний інтелект, 2016, № 2

УДК 004.048

С.А.Бабічев[1], В.І.Литвиненко[2], М.А.Таіф[2], А.О.Фефелов[2]
[1]Університет Яна Евангелиста Пуркіньє в Усті над Лабем, Чехія
вул. Чеської молоді, 8, Усті над Лабем, 40096, Чеська республіка
[2]Херсонський національний технічний університет
Бериславське шосе, 24, Херсон, 73008, Україна

ОЦІНКА ЯКОСТІ ОБРОБКИ СКЛАДНИХ ДАНИХ БІОЛОГІЧНОЇ ПРИРОДИ НА ОСНОВІ КРИТЕРІЇВ ЕНТРОПІЇ

S.A.Babichev[1], V.I.Lytvynenko[2], M.A.Taif[2], A.O.Fefelov[2]
[1] Jan Evangelista Purkyne University in Usti nad Labem, Czech Republic
8, Ceske mladeze Str., Usti nad Labem, 400 96, Czech Republic
[2] The Kherson National Technical University
Beryslavske highway, 24, Kherson, 73008, Ukraine

THE ESTIMATION OF THE COMPLEX BIOLOGICAL DATA PROCESSING BASED ON THE ENTROPY CRITERIA

У статті представлено систему оцінки якості обробки складних даних біологічної природи з використанням критерію ентропії Шеннона. Проведено порівняльний аналіз різних методів розрахунку ентропії Шеннона при використанні модельного сигналу при різних рівнях відношення сигнал-шум. Запропоновано багатокроковий алгоритм обробки даних ДНК мікрочіпів для визначення експресій генів, у якому оцінка якості обробки на кожному етапі здійснюється на основі середнього значення ентропії Шеннона для усіх об’єктів бази даних.
Ключові слова: ентропія Шеннона, експресія генів, ДНК мікрочіп, фільтрація.
The paper presents the system toestimatethe complex biological data qualityprocessing by the Shannon entropy criteria use. The compare analysis of the various methods of the Shannon entropy calculation by the use of the model signals with different levels of noise-to-signal ratiowere carried out during the simulation process. The paper presents also the multi-step algorithm of DNA microarray processing where the estimation of the processing quality at the each step is carried out by the average of the Shannon entropy for all objects of database.
Keywords: Shannon entropy, gene expression, DNA MicroArray, filtration.

Вступ

На сучасному етапі одним із актуальних напрямків біоінформатики є ідентифікація стану біологічних об’єктів шляхом аналізу експресій генів. Даний напрямок пов’язаний зі створенням генних регулюючих мереж, які спрямовані на визначення стану біологічного об’єкта та прогнозування зміни стану з урахуванням характеру взаємодії генів, що характеризують даний об’єкт. На даний час найбільш розповсюдженими є наступні технології визначення експресій генів: технологія мікрочіпів ДНК та технологія секвенування РНК. Кожна з цих технологій має свої недоліки та переваги. Технологія ДНК мікрочіпів є суттєво дешевшою, але отримані вектори експресій генів мають значну шумову складову, що обумовлена процесом створення мікрочіпів та зчитування з них інформації. Технологія секвенування РНК дозволяє отримати вектори експресій генів зі значно меншим відношенням шум-сигнал, але вартість цієї технології значно більша порівняно з вартістю технології ДНК мікрочіпів. Але, у будь якому випадку, отримані вектори експресій генів містять специфічну шумову складову, що обумовлена різноманітним характером протікання біологічних процесів в організмі, які не пов’язані з хворобою, що ідентифікується. Особливістю біологічних даних, що отримані шляхом ДНК мікрочіпів або секвенуванням РНК, є також велика розмірність простору ознак, що ускладнює процес обробки інформації. Тому, одним із актуальних напрямків підвищення об’єктивності прогнозування стану

біологічного об’єкту на основі аналізу експресій генів є створення систем фільтрації даних на попередньому етапі обробки інформації на основі використання сучасних критеріїв оцінки якості даних, одним з яких є критерій ентропії Шеннона.

Постановка проблеми

Структурну схему процесу отримання матриці експресій генів шляхом технології ДНК мікрочіпів представлено на рис. 1.

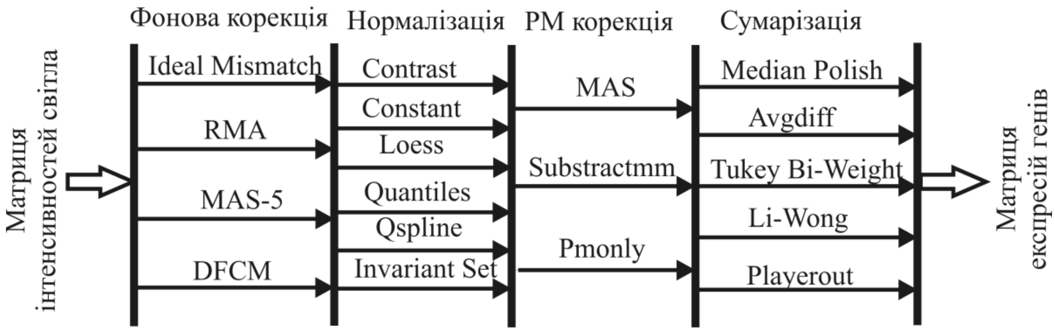


Рис. 1. Структурна схема процесу отримання матриці експресій генів

Як видно з рис. 1, процес трансформації зображення мікрочіпу, яке отримане шляхом лазерного сканування, у матрицю експресій генів складається з чотирьох етапів: фонової корекції, нормалізації даних мікрочіпу, PM корекції та сумаризації. Фонова корекція спрямована на зменшення шуму, що виникає внаслідок процесу сканування мікрочіпу, процес нормалізації дозволяє порівнювати дані, що отримані з різних мікрочіпів при різних умовах проведення експерименту. PM корекція сприяє зменшенню ефекту неспецифічної гібридизації за рахунок урахування MM-проб. На етапі сумаризації визначається експресія відповідного гену шляхом зваженого додавання інтенсивностей світла різних проб, що відповідають даному гену. Кожний етап передбачає використання різних методів, які здійснюють безпосередній вплив на інформативність експресій генів мікрочіпу. Для вибору оптимальної комбінації методів отримання матриці експресій генів необхідне проведення порівняльного аналізу ефективності використання різних комбінацій методів на основі кількісних критеріїв оцінки якості обробки даних. У даній роботі за такий критерій взято ентропію Шеннона, яка визначає кількісну міру невизначеності відповідного стану системи [1,2]:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1),$$

де $p_i = \dfrac{N_i}{N}$ – є ймовірність реалізації *i*-го стану, *N* – об’єм вибірки (кількість станів системи), а N_i – частота повторювання *i*-го стану. При цьому, якщо n – число рівнів дискретизації стану системи, то

$$N = \sum_{i=1}^{n} N_i\,,\ \sum_{i=1}^{n} p_i = 1\,.$$

Слід зазначити, що формула (1) є узагальненою. Її конкретне використання визначається параметром p_i, тобто визначенням простору станів та способом реалізації конкретного стану.

Аналіз останніх досліджень і публікацій

Сьогодні існує велика кількість практичних галузей, де використовуються ентропійні критерії [3-5]. Термін <u>ентропія</u> (грець. *evtropov*– перетворення) уперше ввів

німецький фізик Клаузіус у 1865 році, як міру перетворення теплової енергії у механічну та навпаки. Зв'язок ентропії з інформацією уперше побачив у 1957 році Л. Больцман. Він характеризував ентропію як міру недостатньої інформації про стан системи. Подальші кроки у напрямку розвитку поняття ентропії пов'язані з таким вченими як Гіббс [6], Хартлі [7], Шеннон [1], Колмогоров [8], Рены [9], Тсаллес [10] та фон Нейман [11]. Кількісну адитивну міру для інформації уперше запропонував у 1928 році Хартлі. Згідно з теоремою Хартлі, для знаходження елементу х, який входить до складу множини, що складається з N елементів, необхідна кількість інформації:

$$H = \log_2 N . \quad (2)$$

У загальному випадку N можна вважати кількістю рівноймовірних виходів або статистичною вагою, а H – кількістю інформації для реалізації *i*-го виходу. Шеннон узагальнив формулу Хартлі щодо випадку систем з нерівноймовірнісними станами. У [12] автори представили методику покрокової обробки хроматограм мас спектру наркотичних речовин на основі комплексного використання вейвлет-аналізу та ентропійних критеріїв. Оптимізацію процесу вибору типу вейвлету, рівню вейвлет-декомпозиції, значення трешолдингового коефіцієнту було виконано з використанням ентропій Шеннона та логарифму енергії сигналу. Оптимальний рівень вейвлет-фільтрації було обрано на основі екстремумів відповідних критеріїв. У [13] автором використано поняття ентропійних потенціалів для дослідження різних систем та процесів. Однак, слід зазначити, що, незважаючи на значні успіхи у даний предметній галузі, проблема критеріальної оптимізації вибору методів та засобів обробки складної інформації на даний час не має однозначного рішення.

До невирішеної частини загальної проблеми слід віднести відсутність загальної технології побудови системи попередньої обробки складних даних біологічної природи з метою зменшення шумової складової та розмірності простору ознак, які характеризують об'єкти, що досліджуються, на основі кількісних критеріїв оцінки якості обробки інформації.

Метою роботи є проведення досліджень щодо використання критеріїв ентропії Шеннона для оцінки інформативності даних мікрочіпових експериментів на різних етапах обробки інформації та розробки покрокового алгоритму визначення оптимальної комбінації методів обробки даних, що відповідає максимальній інформативності векторів експресій генів об'єктів, що досліджуються.

Виклад основного матеріалу

Усі методи оцінки ентропії Шеннона можна розділити на дві групи. До першої групи відносяться методи, що засновані на частотах реалізації тієї або іншої події. Згідно з методами другої групи, ентропія розраховується безпосередньо з сигналу без використання вектору частот виникнення відповідних подій. Структурну блок-схему різних методів розрахунку ентропії Шеннона представлено на рис. 2. Відповідно до принципу максимуму ентропії, більш високій ступені упорядкованості інформації, що характеризує об'єкт, відповідає менше значення ентропії Шеннона і навпаки, шумова компонента сигналу має максимальне значення ентропії. При цьому, значення ентропії Шеннона не повинно змінюватися при різних амплітудах шумової компоненти за умови незмінного характеру шуму.

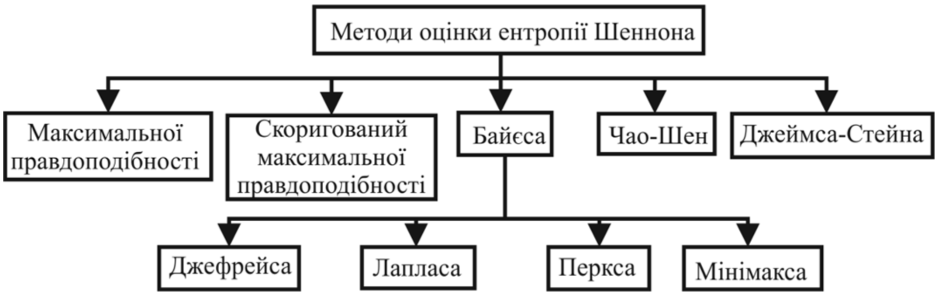


Рис. 2. Структурна блок-схема методів оцінки ентропії Шеннона

Для дослідження якості різних методів оцінки ентропії Шеннона було згенеровано дві групи сигналів. Перша група містила випадкові сигнали при рівні диск ретизації 6000 та різних рівнях амплітуд шумової компоненти. Амплітуда шуму змінювалась від 0,1 до 0,9 умовних одиниць з кроком 0,1. Друга група сигналів відрізнялась від першої значенням амплітуди. Амплітуда шуму сигналів даної групи змінювалась від 50 до 450 умовних одиниць з кроком 50. На рис. 3 та 4 показано графіки зміни значень ентропій Шеннона при різних рівнях шумової компоненти при використанні наведених вище методів розрахунку ентропії. Аналіз характеру зміни значення ентропій при різних рівнях шумової компоненти дозволяє зробити висновок, що при високому рівні шуму значення усіх ентропій лежать у достатньо вузькому діапазоні, а їхню зміну можна пояснити випадковістю сигналу, що досліджується. Вибір критерію розрахунку ентропії Шеннона у цьому випадку не має особливого значення. Інший висновок випливає з аналізу характеру зміні ентропій при низькому рівні шумової компоненти. У даному випадку спостерігається функціональна залежність ентропій Шеннона, що розраховані за методами ММ, Chao-Shen, Jeffreys, Laplace, від рівня шумової компоненти. При зростанні амплітуди шуму дані ентропії зменшувались у більшій або меншій мірі. Спостерігалась хаотична зміна ентропій ML, Perksa та MiniMaxa в досить вузькому діапазоні, що можна пояснити хаотичністю сигналів, що досліджуються. Однак, найвищу стійкість до зміни амплітуди шуму при малих значеннях шумової компоненти показав критерій ентропії, що визначений за методом James-Stein(James-Stein shrinkage estimator). Значення даного критерію не змінювалось протягом зміни амплітуди шуму у заданому діапазоні.

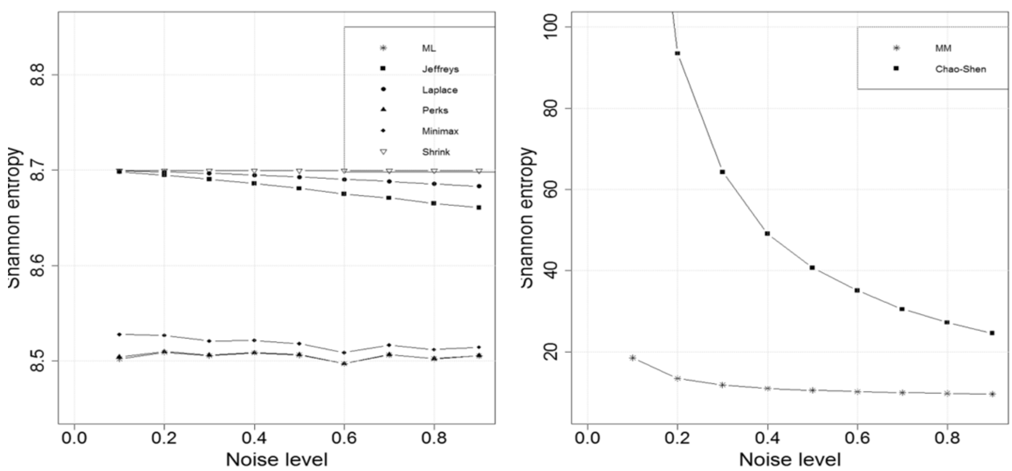


Рис. 3. Графіки ентропій Шеннона при зміні амплітуди шумової компоненти від 0,1 до 0,9 умовних одиниць

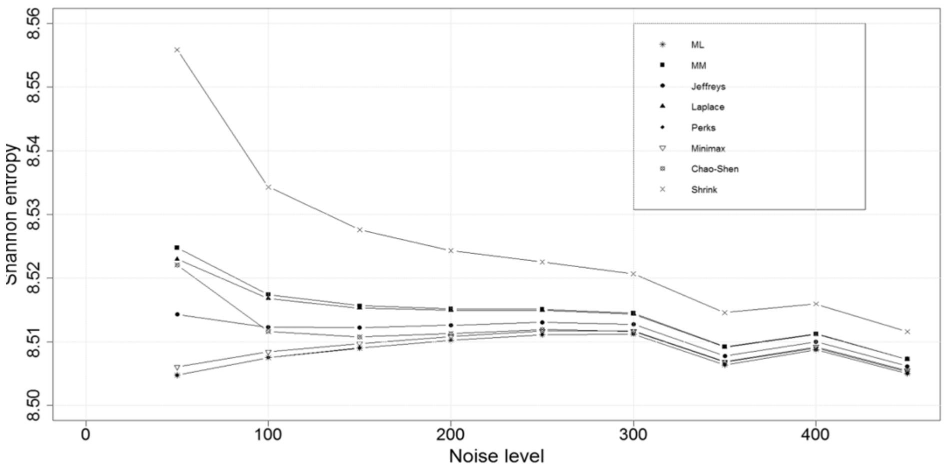


Рис. 4. Графіки ентропій Шеннона при зміні амплітуди шумової компоненти від 50 до 450 умовних одиниць

Оцінку характеру зміни ентропій Шеннона від ступеню зашумленості біологічного сигналу визначимо з використанням даних експресій генів, отриманих шляхом аналізу даних мікрочіпів ДНК хворих на рак легенів GEOD-68571 бази даних Array Express [14], яка включає в себе профілі експресій генів 95 пацієнтів, серед яких 10 є здоровими, а 85 хворих пацієнтів розділені за рівнем розвитку хвороби на три групи: 23 пацієнти мають добрий стан, 41 пацієнт має помірний стан, а 21 пацієнт має поганий стан. Оригінальний сигнал одного з пацієнтів, який був використаний у дослідженнях як базовий, представлений на рис. 5. Даний сигнал являє собою вектор експресій генів клітин органу, що досліджується, при різних умовах визначення експресії. Далі на сигнал накладалася шумова компонента, амплітуда якої змінювалася від 20 до 160 з кроком 20. Даний вибір визначався значеннями експресій генів даних, що досліджуються. Амплітуда шумової компоненти у цьому випадку у багато разів менша за середній рівень експресії генів. На рис. 6 представлений графік зміни ентропій Шеннона при різній мірі зашумленості даних експресій генів.

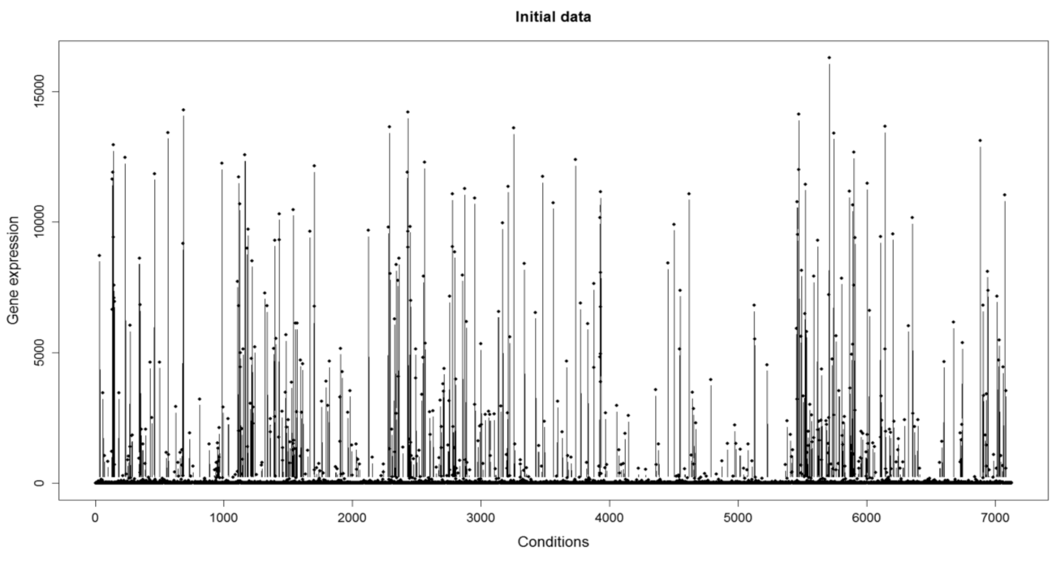


Рис. 5. Оригінальний сигнал експресій генів біологічного об'єкту за різних умов визначення експресії

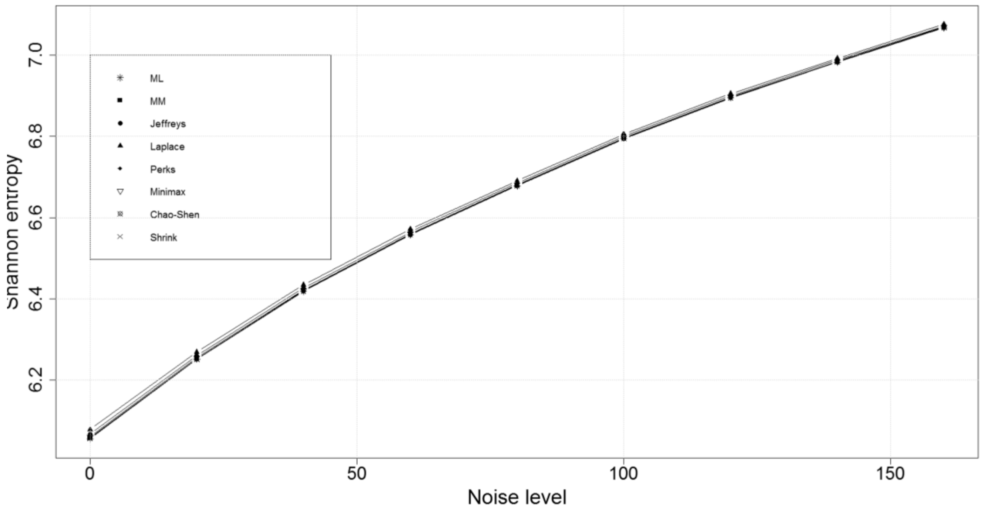


Рис. 6. Графік зміни ентропій Шеннона при різних рівнях зашумленості даних експресій генів

Аналіз графіку підтверджує припущення, що в процесі видалення «білого шуму» ентропія даних експресій генів буде зменшуватись, при цьому значення усіх ентропій змінюється монотонно та погоджено. Аналіз результатів моделювання показує, що для грубої оцінки інформативності біологічного сигналу усі методи оцінки ентропії Шеннона дають аналогічні результати, але для оцінки ентропії шумової компоненти, у процесі її видалення, в системах тонкої очистки метод Джеймса та Стейна (James-Stein shrinkage estimator) має перевагу над іншими методами за рахунок стійкості до зміни амплітуди шуму.

Алгоритм обробки даних мікрочіпів ДНК, з метою підвищення якості процесу визначення експресій генів відповідних об'єктів, представлений на рис. 7. Як критерій об'єктивності, використано середнє значення ентропії Шеннона, що розраховане, з використанням методу Джеймса та Стейна, для усіх мікрочіпів, що досліджуються. Реалізація даного алгоритму передбачає наступні етапи:

Крок 1. Завантаження даних ДНК мікрочіпів у систему обробки інформації.

Крок 2. Завдання етапу обробки даних. Довільна фіксація методів, що не відповідають даному етапу.

Крок 3. Завдання методу обробки даних, що відповідає вибраному етапу. Обробка даних ДНК мікрочіпів даною комбінацією методів.

Крок 4. Розрахунок ентропії Шеннона для векторів експресій генів, що відповідають кожному мікрочіпу, що досліджується. Розрахунок середнього значення ентропії Шеннона для усіх ДНК мікрочіпів.

Крок 5. Якщо порядковий номер методу менший за загальну кількість методів, що відповідають даному етапу, перехід на крок 3. В іншому випадку, вибір та фіксація методу, що відповідає мінімуму середнього значення ентропії Шеннона для усіх масивів даних, що досліджуються.

Крок 6. Якщо порядковий номер етапу менший за максимальну кількість етапів обробки даних, перехід на крок 2. У іншому випадку фіксація остаточного рішення по вибору оптимальної комбінації методів обробки даних для визначення експресій генів мікрочіпів ДНК.

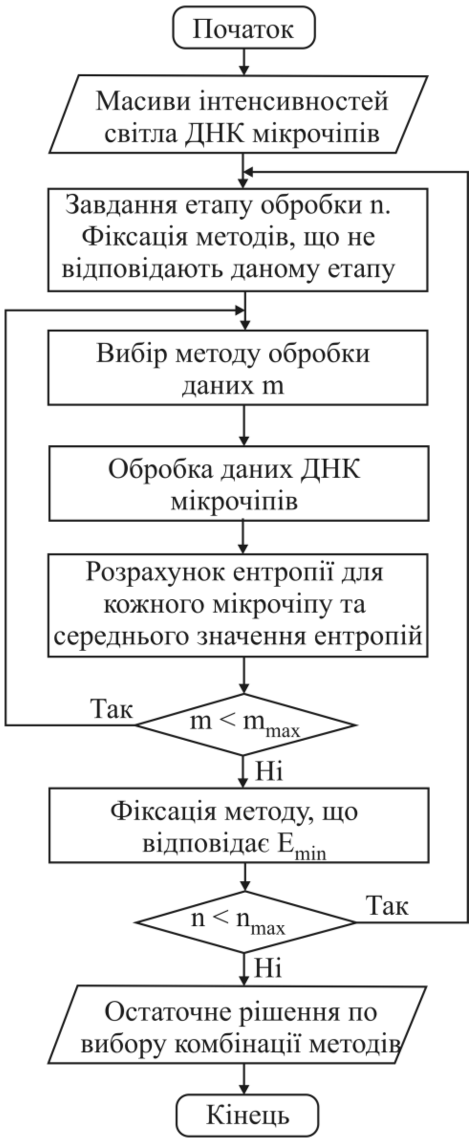


Рис. 7. Алгоритм передобробки даних ДНК мікрочіпів для визначення експресій генів

Моделювання процесу отримання матриці експресій генів з використанням запропонованого алгоритму було проведено з використанням пакету Bioconductor програмного середовища R. На рис. 8а показано діаграму розподілу середнього значення ентропії Шеннона для оригінального зображення та зображень з фоновою корекцією методами «rma», «mass» та «DFCM» відповідно. Метод «IdealMismatch» не використовувався через гіршу якість його роботи (за результатами досліджень компанії Affymetrix) [15].

Аналіз діаграми дозволяє зробити припущення про доцільність використання для даних типів зображень «rma» методу фонової корекції, оскільки значення ентропії Шеннона для об’єктів, що досліджуються, є найменшими порівняно зі значеннями, отриманими при використанні інших методів фонової корекції. Цей факт свідчить про більш високу інформативність даних, отриманих шляхом фонової корекції «rma» методом.

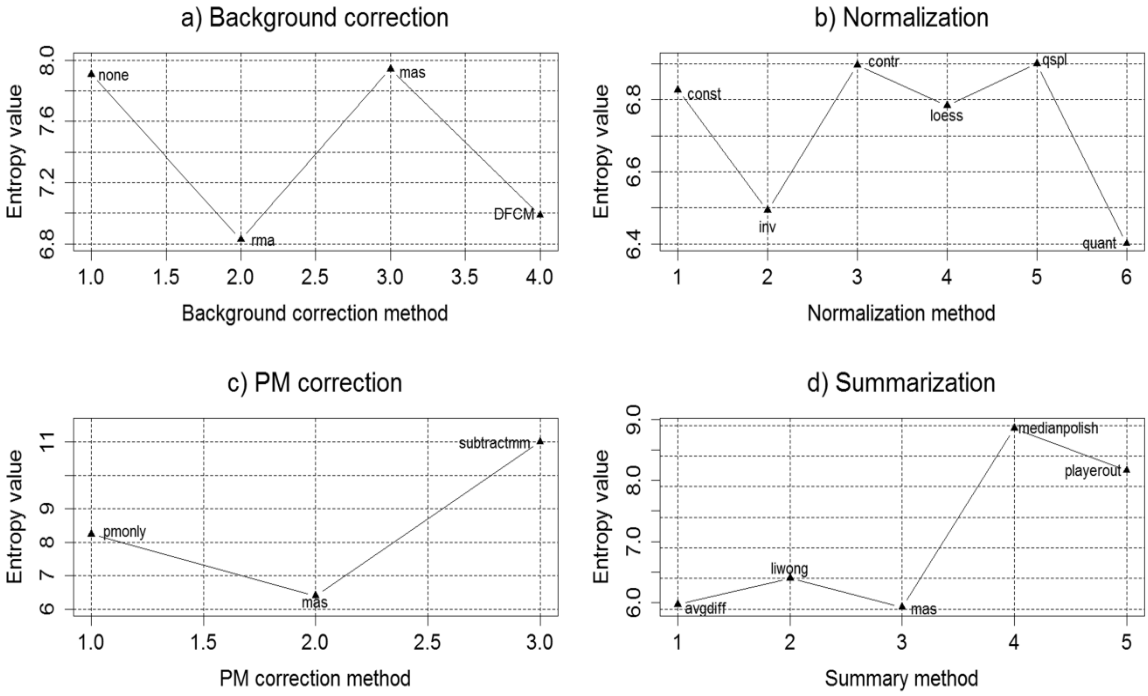


Рис. 8. Діаграми розподілу значень ентропії Шеннона для різних методів обробки даних ДНК мікрочіпів: a) методи фонової корекції; b) методи нормалізації; c) методи PM корекції; d) методи сумаризації

На рис. 8b показані аналогічні діаграми при використанні різних методів нормалізації даних. При цьому, в усіх випадках фонову корекцію було виконано «rma» методом, а PM корекцію та сумаризацію – «mas» та «Li-Wong» методами відповідно. Аналіз діаграм дозволяє зробити висновок, що, з точки зору ентропії Шеннона, найкращим методом для нормалізації даних мікрочіпів є квантильна нормалізація, оскільки середнє значення ентропії при використанні даного методу є також мінімальним. Діаграми розподілу значень ентропії Шеннона при використанні різних методів PM корекції та сумаризації (рис. 1) представлені на рис. 8c та 8d відповідно. Аналіз діаграм на рис. 8c свідчить про доцільність використання «mas» методу PM корекції даних мікрочіпів. Середнє значення ентропії для усіх об'єктів бази даних при використанні даного методу є найменшим. Результати, що представлені на рис. 8d свідчать про доцільність використання «mas» методу для сумаризації інтенсивностей світла проб відповідного гену. Ентропія експресій генів, які характеризують стан відповідного об'єкту, у цьому випадку є мінімальною, що свідчить про більш високу інформативність векторів експресій генів об'єктів, що досліджуються.

На рис. 9 показано діаграми розмаху інтенсивностей світла первинних необроблених даних мікрочіпів (рис. 9a) та експресій генів, отриманих шляхом фонової корекції «rma» методом, квантильної нормалізації, PM корекції та сумаризації «mas» методами (рис. 9b).

Аналіз діаграм підтверджує високу ефективність використання даної комбінації методів. Медіани векторів експресій генів, що відповідають різним мікрочіпам, лежать у дуже вузькому діапазоні (8,24-8,72), при цьому розподіл квантилів відповідних даних дозволяє проводити якісний порівняльний аналіз мікрочіпів, що відповідають різним об'єктам бази даних, що досліджується.

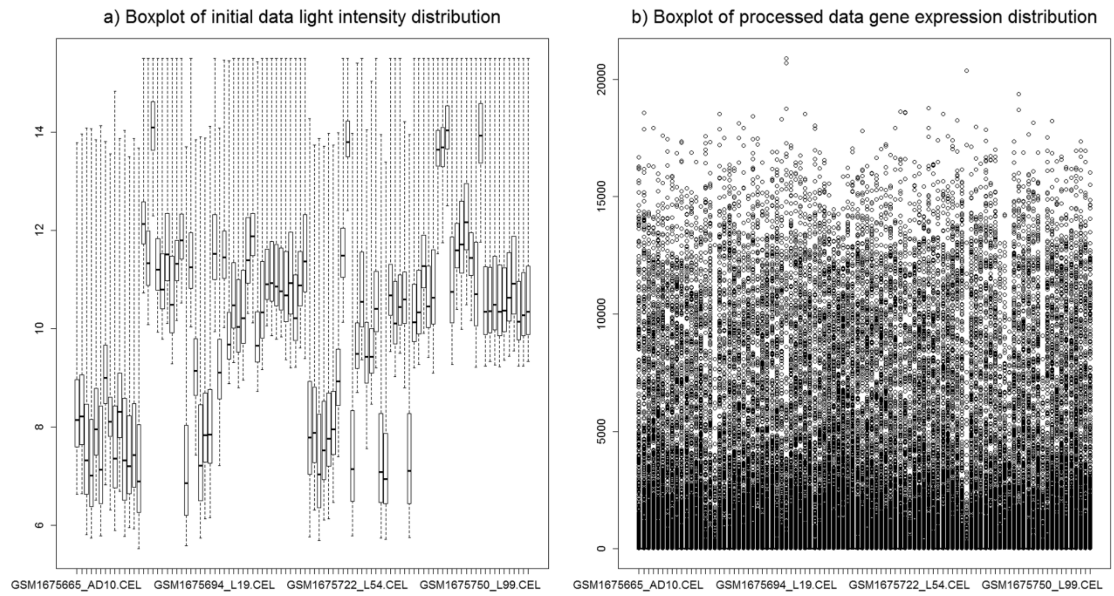


Рис. 9. Діаграми розмаху: a) інтенсивностей світла первинних необроблених даних; b) експресій генів оброблених даних

Висновки

Результати моделювання, що представлені у роботі, підтверджують ефективність використання критеріїв ентропії Шеннона для оцінки якості обробки біологічних даних складної природи. Порівняльний аналіз різних методів розрахунку ентропії Шеннона показав, що для грубої оцінки інформативності даних, що обробляються, усі методи дають аналогічні результати, але на рівні тонкої фільтрації оцінка характеру шумової компоненти, з використанням різних методів розрахунку ентропії Шеннона, є різною. Найбільш стійким до амплітуди шуму при незмінному характері сигналу є метод Джеймса та Стейна (James-Stein shrinkage estimator). Значення даного критерію не змінювалось при зміні амплітуди шуму у рамках заданого діапазону. Це дає можливість створення багатокрокової системи фільтрації складних даних на основі сучасних методів обробки інформації.

У роботі запропоновано алгоритм обробки даних мікрочіпових експериментів для визначення експресій генів об’єктів, що досліджуються. Запропоновано алгоритм визначення оптимальної комбінації методів, які дозволяють отримати вектори експресії генів з більш високою інформативністю, що сприяє підвищенню об’єктивності подальшої ідентифікації об’єктів. Як критерій оцінки якості обробки даних, було використано середнє значення ентропії Шеннона для усіх векторів експресій генів, яке розраховувалося за методом Джеймса та Стейна. Порівняльний аналіз діаграм розмаху оброблених та необроблених даних підтверджує ефективність запропонованої методики. Перспективами подальших досліджень авторів є створення системи фільтрації «білого шуму» та системи редукції простору ознак складних даних біологічної природи на основи критеріїв ентропії Шеннона.

Література

1. Shannon C. E. A mathematical theory of communication.: Bell System Technical Journal. – 1948. – V. 27. – P. 379-423, 623-656.
2. Чумак О.В. Энтропии и фракталы в анализе данных. М.: Ижевск, НИЦ «Регулярная и хаотическая динамика». – 2011. – 164 с.
3. StinsonD.R. Cryptography. Theory and Practice. Chapman&Hall/CRC, 2006. – 611 p.

4. Yeo G. Burge C.M. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals // Computational biology, 2004. – №11. – P. 377-471.
5. Archer E., Park I.M., Pillow J.W. Bayesian Entropy Estimation for Countable Discrete Distribution // Journal of Machine Learning Research, 2014. – P. 2833-2868.
6. Гиббс Дж.В. Термодинамика. Статистическая механика: Избранные труды. М.: Наука, 1982. – 584 с.
7. Хартли Р. В. Л. Передача информации / Теория информации и её приложения / пер. с англ. / под ред. Харкевича. М.: Физматгиз, 1959.
8. Колмогоров А. Н. Об энтропии на единицу времени как метрическом инварианте метаморфизмов. ДАНСССР. – 1959. – Т. 124. – С. 754-755.
9. Renyi A. On measures of entropy and information // Proc. Fourth Berkeley Symposium. Berkeley, Los-Angeles: University of California Press, 1961. – Vol. 1. – P. 547-561.
10. Tsallis C. Possible Generalization of Boltzmann-Gibbs-Statistics // J. Stat. Phys., 1988. – Vol. 52. – P. 479-487.
11. Von Neumann J., Mathematische Grundlagen der Quantenmechanik, Springer, Berlin, 1932.
12. Бабичев С.А., Дидык А.А., Литвиненко В.И., Фефелов А.А., Шкурдода С.В. Фильтрация хроматограмм с помощью вейвлет-анализа с использованием критерия энтропии // Системные технологии. – Днепропетровск, 2011. – № 6(77). – С. 117-131.
13. Лазарев В.Л. Исследование систем на основе энтропийных и информационных характеристик // Журнал технической физики, 2010. – Т. 80, вып. 2. – С. 1-7.
14. Beer D.G., Kardia S.L., Huang C.C., and all. Gene-expression profiles predict survival of patients with lung adenocarcinoma // Nature Medicine,2002. – №8(8). – P. 816-824.
15. Affymetrix. Statistical Algorithms Description Document // Affymetrix, Inc., Santa Clara, CA, 2002. – P. 1-27.

Literatura

1. Shannon C. E. A mathematical theory of communication.: Bell System Technical Journal. – 1948. – V. 27. – P. 379-423, 623-656.
2. Chumak O.V. Jentropiiifraktaly v analizedannyh. M.: Izhevsk, NIC «Reguljarnajaihaoticheskajadinamika». – 2011. – 164 s. (RUS)
3. StinsonD.R. Cryptography. Theory and Practice. Chapman&Hall/CRC, 2006. – 611 p.
4. YeoG. BurgeC.M. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals // Computational biology, 2004. – №11. – P. 377-471.
5. Archer E., Park I.M., Pillow J.W. Bayesian Entropy Estimation for Countable Discrete Distribution // Journal of Machine Learning Research, 2014. – P. 2833-2868.
6. Gibbs Dzh.V. Termodinamika. Statisticheskaja mehanika: Izbrannye trudy. M.: Nauka, 1982. – 584s.
7. Hartli R. V. L. Peredacha informacii / Teorija informacii i ejo prilozhenija / per. s angl. / pod red. Harkevicha. M.: Fizmatgiz, 1959.(RUS)
8. Kolmogorov A. N. Ob jentropii na edinicu vremeni kak metricheskom invariante metamorfizmov. DAN SSSR. – 1959. – T. 124. – S. 754-755.(RUS)
9. Renyi A. On measures of entropy and information // Proc. Fourth Berkeley Symposium. Berkeley, Los-Angeles: University of California Press, 1961. – Vol. 1. – P. 547-561.
10. Tsallis C. Possible Generalization of Boltzmann-Gibbs-Statistics // J. Stat. Phys., 1988. – Vol. 52. – P. 479-487.
11. Von Neumann J., Mathematische Grundlagen der Quantenmechanik, Springer, Berlin, 1932.
12. Babichev S.A., Didyk A.A., Litvinenko V.I., Fefelov A.A., Shkurdoda S.V. Fil'tracijahromatogramm s pomoshh'juvejvlet-analiza s ispol'zovaniemkriterijajentropii // Sistemnyetehnologii. – Dnepropetrovsk, 2011. – № 6(77). – S. 117-131.(RUS)
13. Lazarev V.L. Issledovanie sistemna osnov ejentropijnyhi informacionnyh harakteristik // Zhurnal tehnicheskoj fiziki, 2010. – T. 80, vyp. 2. – S. 1-7.(RUS)
14. Beer D.G., Kardia S.L., Huang C.C., and all. Gene-expression profiles predict survival of patients with lung adenocarcinoma // Nature Medicine,2002. – №8(8). – P. 816-824.
15. Affymetrix. Statistical Algorithms Description Document // Affymetrix, Inc., Santa Clara, CA, 2002. – P. 1-27.

RESUME

S.A. Babichev, V.I. Lytvynenko, M.A. Taif, A.O. Fefelov

The estimation ofthe complex biological data processing based on the entropy criteria

The paper presents the system to estimate the complex biological data quality processing by the Shannon entropy criteria use. As the methods to calculate the Shannon entropy criterion were used follows: maximum likelihood, corrected maximum likelihood, Chao and Shen, James-Stein shrinkage estimator, Jeffreys, Laplace, Perks and minimax. The compare analysis of the various methods of the Shannon entropy calculation by the use of the model signals with different levels of noise-to-signal ratio were carried out during the simulation process. The results of the simulation process show that the best criterion in terms of independence on the level of “white” noise is the James-Stein shrinkage estimator, because the value of this criterion do not change during noise level raise. The data of the biological object gene expression were used to evaluate the change of the Shannon entropy criterion versus the levels of noise-to-signal ratio for complex nature data. The analysis of the simulation results shows that all methods of the Shannon entropy estimation give the same results for primary refining the biological signal information, but the James-Stein shrinkage estimator has the advantage to compare with other methods in case the polishing of the signal. The paper presents also the multi-step algorithm of DNA microarray processing where the estimation of the processing quality at the each step is carried out by the average of the Shannon entropy for all objects of database. The simulation of the process to obtain the gene expression matrix using the proposal algorithm was carried out by the use of package “Bioconductor” of software R. The different methods of the background correction, normalization, PM correction and summarization were studied during the simulation process. The presented technology allows to change an optimal group of methods to increase the informativeness of the obtained data.

Надійшла до редакції 01.11.2016